Case No. 1002

DNA RESTRICTION SITE MAPPING

Field of the Invention

The invention relates generally to methods for construction physical maps of DNA, especially genomic DNA, and more particularly, to a method of providing high resolution physical maps by sequence analysis of concatenations of segments of restriction fragment ends.

BACKGROUND

Physical maps of one or more large pieces of DNA, such as a genome or chromosome, consist of an ordered collection of molecular landmarks that may be used to position, or map, a smaller fragment, such as clone containing a gene of interest, within the larger structure, e.g. U.S. Department of Energy, "Primer on Molecular Genetics," from Human Genome 1991-92 Program Report; and Los Alamos Science, 20: 112-122 (1992). An important goal of the Human Genome Project has been to provide a series of genetic and physical maps of the human genome with increasing resolution, i.e. with reduced distances in basepairs between molecular landmarks, e.g. Murray et al, Science, 265: 2049-2054 (1994); Hudson et al, Science, 270: 1945-1954 (1995); Schuler et al, Science, 274: 540-546 (1996); and so on. Such maps have great value not only in furthering our understanding of genome organization, but also as tools for helping to fill contig gaps in large-scale sequencing projects and as tools for helping to isolate disease-related genes in positional cloning projects, e.g. Rowen et al, pages 167-174, in Adams et al, editors, Automated DNA Sequencing and Analysis (Academic Press, New York, 1994); Collins, Nature Genetics, 9: 347-350 (1995); Rossiter and Caskey, Annals of Surgical Oncology, 2: 14-25 (1995); and Schuler et al (cited above). In both cases, the ability to rapidly construct high-resolution physical maps of large pieces of genomic DNA is highly desirable.

Two important approaches to genomic mapping include the identification and use of sequence tagged sites (STS's), e.g. Olson et al, Science, 245: 1434-1435 (1989); and Green et al, PCR Methods and Applications, 1: 77-90 (1991), and the construction and use of jumping and linking libraries, e.g. Collins et al, Proc. Natl. Acad. Sci., 81: 6812-6816 (1984); and Poustka and Lehrach, Trends in Genetics, 2: 174-179 (1986). The former approach makes maps highly portable and convenient, as maps consist of ordered collections of nucleotide sequences that allow application without having to acquire scarce or specialized reagents and libraries. The latter approach provides a systematic means for identifying molecular landmarks spanning large genetic distances and for ordering such landmarks via hybridization assays with members of a linking library.

Unfortunately, these approaches to mapping genomic DNA are difficult and laborious to implement. It would be highly desirable if there was an approach for constructing physical maps that combined the systematic quality of the jumping and linking libraries with the convenience and portability of the STS approach.

Summary of the Invention

Accordingly, an object of my invention is to provide methods and materials for constructing high resolution physical maps of genomic DNA.

Another object of my invention is to provide a method of ordering restriction fragments from multiple enzyme digests by aligning matching sequences of their ends.

Still another object of my invention is to provide a high resolution physical map of a target polynucleotide that permits directed sequencing of the target polynucleotide with the sequences of the map.

Another object of my invention is to provide vectors for excising ends of restriction fragments for concatenation and sequencing.

Still another object of my invent is to provide a method monitoring the expression of genes.

A further object of my invention is to provide physical maps of genomic DNA that consist of an ordered collection of nucleotide sequences spaced at an average distance of a few hundred to a few thousand bases.

My invention achieves these and other objects by providing methods and materials for determining the nucleotide sequences of both ends of restriction fragments obtained from multiple enzymatic digests of a target polynucleotide, such as a fragment of a genome, or chromosome, or an insert of a cosmid, BAC, YAC, or the like. In accordance with the invention, a polynucleotide is separately digested with different combinations of restriction endonucleases and the ends of the restriction fragments are sequenced so that pairs of sequences from each fragment are produced. A physical map of the polynucleotide is constructed by ordering the pairs of sequences by matching the identical sequences among such pairs resulting from all of the digestions.

In the preferred embodiment, a polynucleotide is mapped by the following steps: (a) providing a plurality of populations of restriction fragments, the restriction fragments of each population having ends defined by digesting the polynucleotide with a plurality of combinations of restriction endonucleases; (b) determining the nucleotide sequence of a portion of each end of each restriction fragment of each population so that a pair of nucleotide sequences is obtained for each restriction fragment of each population; and (c) ordering the pairs of nucleotide sequences by matching the nucleotide sequences between pairs to form a map of the polynucleotide.

Another aspect of the invention is the monitoring gene expression by providing pairs of segments excised from cDNAs. In this embodiment, segments from each end of each cDNA of a population of cDNAs are ligated together to form pairs, which serve to identify their associated cDNAs. Concatenations of such pairs are sequenced by conventional techniques to provide information on the relative frequencies of expression in the population.

The invention provides a means for generating a high density physical map of target polynucleotides based on the positions of the restriction sites of predetermined restriction endonucleases. Such physical maps provide many advantages, including a more efficient means for directed sequencing of large DNA fragments, the positioning of expression sequence tags and cDNA sequences on large genomic fragments, such as BAC library inserts, thereby making positional candidate mapping easier; and the like.

Brief Description of the Drawings

Figure 1 graphically illustrates the concept of a preferred embodiment of the invention.

Figure 2 provides a diagram of a vector for forming pairs of nucleotide sequences in accordance with a preferred embodiment of the invention.

Figure 3 illustrates a scheme for carrying out the steps of a preferred embodiment of the invention.

Figure 4 illustrates locations on yeast chromosome 1 where sequence information is provided in a physical map based on digestions with Hind III, Eco RI, and Xba I in accordance with the invention.

Definitions

As used herein, the process of "mapping" a polynucleotide means providing a ordering, or series, of sequenced segments of the polynucleotide that correspond to the actual ordering of the segments in the polynucleotide. For example, the following set of five-base sequences is a map of the polynucleotide below (SEQ ID NO: 1), which has the ordered set of sequences making up the map underlined:

(gggtc, ttatt, aacct, catta, ccgga)

GTTGGGTCAACAAATTACCTTATTGTAACCTTCGCATTAGCCGGAGCCT

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides, and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units, e.g. 40-60. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. Usually oligonucleotides comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of polynucleotide present in the population.

DETAILED DESCRIPTION OF THE INVENTION

In accordance with the present invention, segments of nucleotides at each end of restriction fragments produced from multiple digestions of a polynucleotide are sequenced and used to arrange the fragments into a physical map. Such a physical map consists of an ordered collection of the nucleotide sequences of the segments immediately adjacent to the cleavage sites of the endonucleases used in the digestions. Preferably, after each digestion, segments are removed from the ends of each restriction fragment by cleavage with a type IIs restriction endonuclease. Excised segments from the same fragment are ligated together to form a pair of segments. Preferably, collections of such pairs are concatenated by ligation, cloned, and sequenced using conventional techniques.

The concept of the invention is illustrated in Figure 1 for an embodiment which employs three restriction endonucleases: r, q, and s. Polynucleotide (50) has recognition sites (r_1, r_2, r_3, and r_4) for restriction endonucleases r, recognition sites (q_1 through q_4) for restriction endonuclease q, and recognition sites (s_1 through s_5) for restriction endonuclease s. In accordance with the preferred embodiment, polynucleotide (50) is separately digested with r and s, q and s, and r and q to produce three populations of restriction fragments (58), (60), and (62), respectively. Segments adjacent to the ends of each restriction fragment are sequenced to form sets of pairs (52), (54), and (56) of nucleotide sequences, which for sake of illustration are shown directly beneath their corresponding restriction fragments in the correct order. Pairs of sequences from all three sets are ordered by matching sequences between pairs as shown (70). A nucleotide sequence (72) from a first pair is matched with a sequence (74) of a second pair whose other sequence (76), in turn, is matched with a sequence (78) of a third pair. The matching continues, as (80) is matched with (82), (84) with (86), (88) with (90), and so on, until the maximum number of pairs are included. It is noted that some pairs (92) do not contribute to the map. These correspond to fragments having the same restriction site at both ends. In other word, they correspond to situations where there are two (or more) consecutive restriction sites of the same type without other sites in between, e.g. s_3 and s_4 in this example. Preferably, algorithms used for assembling a physical map from the pairs of sequences can eliminate pairs having identical sequences.

Generally, a plurality of enzymes is employed in each digestion. Preferably, at least three distinct recognition sites are used. This can be accomplished by using three or more restriction endonucleases, such as Hind III, Eco RI, and Xba I, which recognize different nucleotide sequences, or by using restriction endonucleases recognizing the same nucleotide sequence, but which have different methylation sensitivities. That is, it is understood that a different "recognition site" may be different solely by virtue of a different methylation state. Preferably, a set of at least three recognition endonucleases is employed in the method of the invention. From this set a plurality of combinations of restriction endonucleases is formed for separate digestion of a target polynucleoitde. Preferably, the combinations are "n-1" combinations of the set. In other words, for a set of n restriction endonucleases, the preferred combinations are all the combinations of n-1 restriction endonucleases. For example, as illustrated in Figure 1 where a set of three restriction endonucleases (r, q, and s) are employed, the n-1 combinations are (r, q), (r, s), and (q, s). Likewise, if four restriction endonucleases (r, q, s, and w) are employed, the n-1 combinations are (r, q, s), (r, q, w), (r, s, w), and (q, s, w). It is readily seen that where a set of n restriction endonucleases are employed the plurality of n-1 combinations is n.

Preferably, the method of the invention is carried out using a vector, such as that illustrated in Figure 2. The vector is readily constructed from commercially available materials using conventional recombinant DNA techniques, e.g. as disclosed in Sambrook et al, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989). Preferably, pUC-based plasmids, such as pUC19, or λ-based phages, such as λ ZAP Express (Stratagene Cloning Systems, La Jolla, CA), or like vectors are employed. Important features of the vector are recognition sites (204) and (212) for two type IIs restriction endonucleases that flank restriction fragment (208). For convenience, the two type IIs restriction enzymes are referred to herein as "IIs_1" and "IIs_2", respectively. IIs_1 and IIs_2 may be the same or different. Recognition sites (204) and (212) are oriented so that the cleavage sites of IIs_1 and IIs_2 are located in the interior of restriction fragment (208). In other words, taking the 5' direction as "upstream" and the 3' direction as "downstream," the cleavage site of IIs_1 is downstream of its recognition site and the cleavage site of IIs_2 is upstream of its recognition site. Thus, when the vector is cleaved with IIs_1 and IIs_2 two segments (218) and (220) of restriction fragment (208) remain attached to the vector. The vector is then re-circularized by ligating the two ends together, thereby forming a pair of segments. If such cleavage results in one or more single stranded overhangs, i.e. one or more non-blunt ends, then the ends are preferably rendered blunt prior to re-circularization, for example, by digesting the protruding strand with a nuclease such as Mung bean nuclease, or by extending a 3' recessed strand, if one is produced in the digestion. The ligation reaction for re-circularization is carried out under conditions that favor the formation of covalent circles rather than concatemers of the vector. Preferably, the vector concentration for the ligation is between about 0.4 and about 4.0 μg/ml of vector DNA, e.g. as disclosed in Collins et al, Proc. Natl. Acad. Sci., 81: 6812-6812 (1984), for λ-based vectors. For vectors of different molecular weight, the concentration range is adjusted appropriately.

In the preferred embodiments, the number of nucleotides identified depends on the "reach" of the type IIs restriction endonucleases employed. "Reach" is the amount of separation between a recognition site of a type IIs restriction endonuclease and its cleavage site, e.g. Brenner, U.S. patent 5,559,675. The conventional measure of reach is given as a ratio of integers, such as "(16/14)", where the numerator is the number of nucleotides from the recognition site in the 5'→3' direction that cleavage of one strand occurs and the denominator is the number of nucleotides from the recognition site in the 3'→5' direction that cleavage of the other strand occurs. Preferred type IIs restriction endonucleases for use as IIs_1 and IIs_2 in the preferred embodiment include the following: Bbv I, Bce 83 I, Bcef I, Bpm I, Bsg I, BspLU 11 III, Bst 71 I, Eco 57 I, Fok I, Gsu I, Hga I, Mme I, and the like. In the preferred embodiment, a vector is selected which does not contain a recognition site, other than (204) and (212), for the type IIs enzyme(s) used to generate pairs of segments; otherwise, re-circularization cannot be carried out.

Preferably, a type IIs restriction endonuclease for generating pairs of segments has as great a reach as possible to maximize the probability that the nucleotide sequences of the segments are unique. This in turn maximizes the probability that a unique physical map can be assembled. If the target polynucleotide is a bacterial genome of 1 megabase, for a restriction endonuclease with a six basepair recognition site, about 250 fragments are generated (or about 500 ends) and the number of nucleotides determined could be as low as five or six, and still have a significant probability that each end sequence would be unique. Preferably, for polynucleotides less than or equal to 10 megabases, at least 8 nucleotides are determined in the regions adjacent to restriction sites, when a restriction endonuclease having a six basepair recognition site is employed. Generally for polynucleotides less than or equal to 10 megabases, 9-12 nucleotides are preferably determined to ensure that the end sequences are unique. In the preferred embodiment, type IIs enzymes having a (16/14) reach effectively provide 9 bases of unique sequence (since blunting reduces the number of bases to 14 and 5 bases are part of the recognition sites (206) or (210)). In a polynucleotide having a random sequence of nucleotides, a 9-mer appears on average about once every 262,000 bases. Thus, 9-mer sequences are quite suitable for uniquely labeling restriction fragments of a target polynucleotide corresponding to a typical yeast artificial chromosome (YACs) insert, i.e. 100-1000 kilobases, bacterial artificial chromosome (BAC) insert, i.e. 50-250 kilobases, and the like.

Immediately adjacent to IIs sites (204) and (212) are restriction sites (206) and (210), respectively that permit restriction fragment (208) to be inserted into the vector. That is, restriction site (206) is immediately downstream of (204) and (210) is immediately upstream of (212). Preferably, sites (204) and (206) are as close together as possible, even overlapping, provided type IIs site (206) is not destroyed upon cleavage with the enzymes for inserting restriction fragment (208). This is desirable because the recognition site of the restriction endonuclease used for generating the fragments occurs between the recognition site and cleavage site of type IIs enzyme used to remove a segment for sequencing, i.e. it occurs within the "reach" of the type IIs enzyme. Thus, the closer the recognition sites, the larger the piece of unique sequence can be removed from the fragment. The same of course holds for restriction sites (210) and (212). Preferably, whenever the vector employed is based on a pUC plasmid, restriction sites (206) and (210) are selected from either the restriction sites of polylinker region of the pUC plasmid or from the set of sites which do not appeal in the pUC. Such sites include Eco RI, Apo I, Ban II, Sac I, Kpn I, Acc65 I, Ava I, Xma I, Sma I, Bam HI, Xba I, Sal I, Hinc II, Acc I, BspMI, Pst I, Sse8387 I, Sph I, Hind III, Afl II, Age I, Bsp120 I, Asc I, Bbs I, Bcl I, Bgl II, Blp I, BsaA I, Bsa BI, Bse RI, Bsm I, Cla I, Bsp EI, BssH II, Bst BI, BstXI, Dra III, Eag I, Eco RV, Fse I, Hpa I, Mfe I, Nae I, Nco I, Nhe I, Not I, Nru I, Pac I, Xho I, Pme I, Sac II, Spe I, Stu I, and the like. Preferably, six-nucleotide recognition sites (i.e. "6-cutters") are used, and more preferably, 6-cutters leaving four-nucleotide protruding strands are used.

Preferably, the vectors contain primer binding sites (200) and (216) for primers p_1 and p_2, respectively, which may be used to amplify the pair of segments by PCR after re-circularization. Recognition sites (202) and (214) are for restriction endonucleases w_1 and w_2, which are used to cleave the pair of segments from the vector after amplification. Preferably, w_1 and w_2, which may be the same or different, are type IIs restriction endonucleases whose cleavage sites correspond to those of (206) and (210), thereby removing surplus, or non-informative, sequence (such as the recognition sites (204) and (212)) and generating protruding ends that permit concatenation of the pairs of segments.

Figure 3 illustrates steps in a preferred method using vectors of Figure 2. Genomic or other DNA (400) is obtained using conventional techniques, e.g. Herrmann and Frischauf, Methods in Enzymology, 152: 180-183 (1987); Frischauf, Methods in Enzymology, 152: 183-199 (1987), or the like, after which it is divided (302) into aliquots that are separately digested (310) with combinations restriction endonucleases, as shown in Figure 3 for the n-1 combinations of the set of enzymes r, s, and q. Preferably, the resulting fragments are treated with a phosphatase to prevent ligation of the genomic fragments with one another before or during insertion into a vector. Restriction fragments are inserted (312) into vectors designed with cloning sites to specifically accept the fragments. That is, fragments digested with r and s are inserted into a vector that accepts r-s fragments. Fragments having the same ends, e.g. r-r and s-s, are not cloned since information derived from them does not contribute to the map. r-s fragments are of course inserted into the vector in both orientations. Thus, for a set of three restriction endonucleases, only three vectors are required, e.g. one each for accepting r-s, r-q, and s-q fragments. Likewise, for a set of four restriction endonucleases, e.g. r, s, q, and t, only six vectors are required, one each for accepting r-s, r-q, r-t, s-q, s-t, and q-t fragments.

After insertion, a suitable host is transformed with the vectors and cultured, i.e. expanded (314), using conventional techniques. Transformed host cells are then selected, e.g. by plating and picking colonies using a standard marker, e.g. β-glactosidase/X-gal. A large enough sample of transformed host cells is taken to ensure that every restriction fragment is present for analysis with a reasonably large probability. This is similar to the problem of ensuring representation of a clone of a rare mRNA in a cDNA library, as discussed in Sambrook et al, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989), and like references. Briefly, the number of fragments, N, that must be in a sample to achieve a given probability, P, of including a given fragment is the following: $N=\ln(1-P)/\ln(1-f)$, where f is the frequency of the fragment in the population. Thus, for a population of 500 restriction fragments, a sample containing 3454 vectors will include at least one copy of each fragment (i.e. a complete set) with a probability of 99.9%; and a sample containing 2300 vectors will include at least one copy of each fragment with a probability of 99%. The table below provides the results of similar calculations for target polynucleotides of different sizes:

Table I

Size of Target Polynucleotide (basepairs)	Average fragment size after cleavage with 2 six-cutters (No. of fragments) [Sample size for complete set with 99% probability]	Average fragment size after cleavage with 3 six-cutters (No. of fragments) [Sample size for complete set with 99% probability]
2.5×10^5	2048 (124) [576]	1365 (250) [1050]
5×10^5	2048 (250) [1050]	1365 (500) [2300]
1×10^6	2048 (500) [2300]	1365 (1000) [4605]

After selection, the vector-containing hosts are combined and expanded in cultured. The vectors are then isolated, e.g. by a conventional mini-prep, or the like, and cleaved with IIs_1 and IIs_2 (316). The fragments comprising the vector and ends (i.e. segments) of the restriction fragment insert are isolated, e.g. by gel electrophoresis, blunted (316), and re-circularized (320). The resulting pairs of segments in the re-circularized vectors are then amplified (322), e.g. by polymerase chain reaction (PCR), after which the amplified pairs are cleaved with w (324) to free the pairs of segments, which are isolated (326), e.g. by gel electrophoresis. The isolated pairs are concatenated (328) in a conventional ligation reaction to produce concatemers of various sizes, which are separated, e.g. by gel electrophoresis. Concatemers greater than about 200-300 basepairs are isolated and cloned (330) into a standard sequencing vector, such as M13. The sequences of the cloned concatenated pairs are analyzed on a conventional DNA sequencer, such as a model 377 DNA sequencer from Perkin-Elmer Applied Biosystems Division (Foster City, CA).

In the above embodiment, the sequences of the pairs of segments are readily identified between sequences for the recognition site of the enzymes used in the digestions. For example, when pairs are concatenated from fragments of the r and s digestion after cleavage with a type IIs restriction endonuclease of reach (16/14), the following pattern is observed (SEQ ID NO: 1):

```
... NNNNrrrrrNNNNNNNNNNNNNNNNNNqqqqqNNNNNN ...
```

where "r" and "q" represent the nucleotides of the recognition sites of restriction endonuclease r and q, respectively, and where the N's are the nucleotides of the pairs of segments. Thus, the pairs are recognized by their length and their spacing between known recognition sites.

Pairs of segments are ordered by matching the sequences of segments between pairs. That is, a candidate map is built by selecting pairs that have one identical and one different sequence. The identical sequences are matched to form a candidate map, or ordering, as illustrated below for pairs (s_1, s_2), (s_3, s_2), (s_3, s_4), (s_5, s_4), and (s_5, s_6), where the "s_k's" represent the nucleotide sequences of the segments:

```
        ...     s1---- s2
     s3---------------- s2
     s3-------------------------------   --- s4
                                  s5------ s4
                                  s5---------------- s6  ...
```

Sequence matching and candidate map construction is readily carried out by computer algorithms, such as the Fortran code provided in Appendix A. Preferably, a map construction algorithm initially sorts the pairs to remove identical pairs prior to map construction. That is, preferably only one pair of each kind is used in the reconstruction. If for two pairs, (s_i, s_j) and (s_m, s_n), $s_i=s_m$ and $s_j=s_n$, then one of the two can be eliminated prior to map construction. As pointed out above, such additional pairs either correspond to restriction fragments such as (92) of Figure 1 (no sites of a second or third restriction endonuclease in its interior) or they are additional copies of pairs (because of sampling) that can be used in the analysis. Preferably, an algorithm selects the largest candidate map as a solution, i.e. the candidate map that uses the maximal number of pairs.

The vector of Figure 2 can also be used for determining the frequency of expression of particular cDNAs in a cDNA library. Preferably, cDNAs whose frequencies are to be determined are cloned into a vector by way of flanking restriction sites that correspond to those of (206) and (210). Thus, cDNAs may be cleaved from the library vectors and directionally inserted into the vector of Figure 2. After insertion, analysis is carried out as described for the mapping embodiment, except that a larger number of concatemers are sequenced in order to obtain a large enough sample of cDNAs for reliable data on frequencies.

Example 1

Constructing a Physical Map of Yeast Chromosome 1
with Hind III, Eco RI, and Xba I

In this example, a physical map of the 230 kilobase yeast chromosome 1 is constructed using pUC19 plasmids modified in accordance with Figure 2. The chromosome is separately digested to completion with the following combinations of enzymes: Hind III and Eco RI, Hind III and Xba I, and Eco RI and Xba I to generate three populations of restriction fragments. Fragments from each population are inserted into separate pUC19 plasmids, one for each restriction fragment having different ends. That is, restriction fragments from the Hind III-Eco RI digestion are present in three types, ones with a Hind III-digested end and an Eco RI-digested end ("H-E" fragments), one with only Hind III-digested ends ("H-H" fragments), and ones with only Eco RI-digested fragments ("E-E" fragments). Likewise, restriction fragments from the Hind III-Xba I digestion are present in three types, ones with a Hind III-digested end and an Xba I-digested end ("H-X" fragments), one with only Hind III-digested ends ("H-H" fragments), and ones with only Xba I-digested fragments ("X-X" fragments). Finally, restriction fragments from the Xba I-Eco RI digestion are present in three types, ones with a Xba I-digested end and an Eco RI-digested end ("X-E" fragments), one with only Xba I-digested ends ("X-X" fragments), and ones with only Eco RI-digested fragments ("E-E" fragments). Thus, the plasmid for the Hind III-Eco RI digestion accepts H-E fragments; the plasmid for the Hind III-Xba I digestion accepts H-X fragments; and the plasmid for the Xba I-Eco RI digestion accepts X-E fragments. The construction of the plasmid for accepting H-E fragments is described below. The other plasmids are construction in a similar manner. Synthetic oligonucleotides (i) through (iv) are combined with a Eco I- and Hind III-digested pUC19 in a ligation reaction so that they assemble into the double stranded insert of Formula I.

```
(i)     5'-AATTAGCCGTACCTGCAGCAGTGCAGAAGCTTGCGT        (SEQ ID NO: 2)

(ii)    5'-AAACCTCAGAATTCCTGCACAGCTGCGAATCATTCG        (SEQ ID NO: 3)

(iii)   5'-AGCTCGAATGATTCGCAGCTGTGCAGGAATTCTGAG        (SEQ ID NO: 4)

(iv)    5'-GTTTACGCAAGCTTCTGCACTGCTGCAGGTACGGCT        (SEQ ID NO: 5)
```

```
                       →      →
                     Bbv I  Bsg I  Hind III
                       ↓      ↓      ↓
5'-AATTAGCCGTACCTGCAGCAGTGCAGAAGCTTGCGTAAACCTCA-
       TCGGCATGGACGTCGTCACGTCTTCGAACGCATTTGGAGT-
         p1 primer binding site


                      p2 primer binding site
         -GAATTCCTGCACAGCTGCGAATCATTCG
         -CTTAAGGACGTGTCGACGCTTAGTAAGCTCGA
              ↑      ↑      ↑
           Eco RI  Bsg I  Bbv I
                     ←      ←
```

Formula I (SEQ ID NO: 6)

Note that the insert has compatible ends to the Eco RI-Hind III-digested plasmid, but that the original Eco RI and Hind III sites are destroyed upon ligation. The horizontal arrows above and below the Bsg I and Bbv I sites indicate the direction of the cleavage site relative to the recognition site of the enzymes. After ligation, transformation of a suitable host, and expansion, the modified pUC19 is isolated and the insert is sequenced to confirm its identity.

Yeast chromosome 1 DNA is separated into three aliquots of about 5 µg DNA (.033 pmol) each, which are then separately digested to completion with Hind III and Eco RI, Hind III and Xba I, and Eco RI and Xba I, respectively. For each of the three populations, the same procedure is followed, which is described as follows for the pUC19 designed for H-E fragments.

Since each enzyme recognizes a six basepair recognition sequence, about 100-140 fragments are produced for a total of about 3.3 pmol of fragments, about fifty percent of which are H-E fragments. 5.26 µg (3 pmol) of plasmid DNA is digested with Eco RI and Hind III in Eco RI buffer as recommended by the manufacturer (New England Biolabs, Beverly, MA), purified by phenol extraction and ethanol precipitation, and ligated to the H-E fragments of the mixture in a standard ligation reaction. A bacterial host is transformed, e.g. by electroporation, and plated so that hosts containing recombinant plasmids are identified by white colonies. The digestion of the yeast chromosome 1 generates about 124 fragments of the three types, about fifty percent of which are H-E fragments and about twenty-five percent each are H-H or E-E fragments. About 290 colonies are picked for H-E fragments, and about 145 each are picked for H-H and E-E fragments. The same procedure is carried out for all the other types of fragments, so that six populations of transformed hosts are obtained, one each for H-E, H-X, X-E, H-H, E-E, and X-X fragments. Each of the populations is treated separately as follows: About 10 µg of plasmid DNA is digested to completion with Bsg I using the manufacturer's protocol (New England Biolabs, Beverly, MA) and after phenol extraction the vector/segment-containing fragment is isolated, e.g. by gel electrophoresis. The ends of the isolated fragment are then blunted by Mung bean nuclease (using the manufacturer's recommended protocol, New England Biolabs), after which the blunted fragments are purified by phenol extraction and ethanol precipitation. The fragments are then resuspended in a ligation buffer at a concentration of about .05 µg/ml in 20 1-ml reaction volumes. The dilution is designed to promote self-ligation of the fragments, following the protocol of Collins et al, Proc. Natl. Acad. Sci., 81: 6812-6816 (1984). After ligation and concentration by ethanol precipitation, phages from the 20 reactions are combined. The pairs of segments carried by the plasmids are then amplified by PCR using primers p_1 and p_2. The amplified product is purified by phenol extraction and ethanol precipitation, after which it is cleaved with Bbv I using the manufacturer's recommended protocol (New England Biolabs). After isolation by polyacrylamide gel electrophoresis, the pairs are concatenated by carrying out a conventional ligation reaction. The concatenated fragments are then separated by polyacrylamide gel electrophoresis and concatemers greater than about 200 basepairs are isolated and ligated into an equimolar mixture of three Phagescript SK sequencing vectors (Stratagene Cloning Systems, La Jolla, CA), separately digested with Hind III, Eco RI, and Hind III and Eco RI, respectively. (Other appropriate mixtures and digestions are employed when different combinations of enzymes are used). Preferably, a number of clones are expanded and sequenced that ensure with a probability of at least 99% that all of the pairs of the aliquot are sequenced. A "lane" of sequence data (about 600 bases) obtained with conventional sequencing provides the sequences of about 25 pairs of segments. Thus, after transfection, about 13 individual clones are expanded and sequenced on a commercially available DNA sequencer, e.g. PE Applied Biosystems model 377, to give the identities of about 325 pairs of segments. The other sets of fragments require an additional 26 lanes of sequencing (13 each for the H-X and X-E fragments).

Figure 4 illustrates the positions on yeast chromosome 1 of pairs of segments ordered in accordance with the algorithm of Appendix A. The relative spacing of the segments along the chromosome is only provided to show the distribution of sequence information along the chromosome.

Example 2

Directed Sequencing of Yeast Chromosome 1 Using Restriction Map Sequences as Spaced PCR Primers

In this example, the 14-mer segments making up the physical map of Example 1 are used to separately amplify by PCR fragments that collectively cover yeast 1 chromosome. The PCR products are inserted into standard M13mp19, or like, sequencing vectors and sequenced in both the forward and reverse directions using conventional protocols. For fragments greater than about 800 basepairs, the sequence information obtained in the first round of sequencing is used to synthesized new sets of primers for the next round of sequencing. Such directed sequencing continues until each fragment is completely sequenced. Based on the map of Example 1, 174 primers are synthesized for 173 PCRs. The total number of sequencing reactions required to cover yeast chromosome 1 depends on the distribution of fragment sizes, and particularly, how many rounds of sequencing are required to cover each fragment: the larger the fragment, the more rounds of sequencing that are required for full coverage. Full coverage of a fragment is obtained when inspection of the sequence information shows that complementary sequences are being identified. Below, it is assumed that conventional sequencing will produce about 400 bases at each end of a fragment in each round. Inspection shows that the distribution of fragment sizes from the Example 1 map of yeast chromosome 1 are shown below together with reaction and primer requirements:

Round of Sequencing	Fragment size range	Number of Fragments	Number of Seq. or PCR Primers	Number of Sequencing Reactions
1	>0	174	174	348
2	>800	92	184	184
3	>1600	53	106	106
4	>2400	28	56	56
5	>3200	16	32	32
6	>4000	7	14	14
7	>4800	5	10	10
8	>5600	1	2	2
		Total No. of Primers:	578	752
			Seq. reactions for map:	39
			Total No. of Reactions:	791

This compares to about 2500-3000 sequencing reactions that are required for full coverage using shotgun sequencing.

Appendix A

Computer Code for Ordering Pairs into a Physical Map

```
      program opsort
c
c     opsort reads ordered pairs from disk files
c     p1.dat, p2.dat, and p3.dat. and sorts
c     them into a physical map.
c
      character*1 op(1000,2,14),w(14),x(14)
      character*1 fp(1000,2,14),test(14)
c
c
      open(1,file='p1.dat',status='old')
      open(5,file='olist.dat',status='replace')
c
c
      nop=0
      read(1,100)nop1
      nop=nop + nop1
      do 101 j=1,nop
        read(1,102)(w(i),i=1,14),
    +             (x(k),k=1,14)
      do 121 kk=1,14
        op(j,1,kk)=w(kk)
        op(j,2,kk)=x(kk)
 121    continue
 101  continue
      read(1,100)nop2
      nop=nop + nop2
      do 1011 j=nop1+1,nop
        read(1,102)(w(i),i=1,14),
    +             (x(k),k=1,14)
      do 1211 kk=1,14
        op(j,1,kk)=w(kk)
        op(j,2,kk)=x(kk)
 1211     continue
 1011  continue
 c
      close(1)
 c
      write(5,110)nop1,nop2,nop
 110  format(3(2x,i4))
 c
 c
      open(1,file='p2.dat',status='old')
      read(1,100)nop3
      nop=nop + nop3
      do 104 j=nop1+nop2+1,nop
        read(1,102)(w(i),i=1,14),
    +             (x(k),k=1,14)
```

```
          do 122 kk=1,14
             op(j,1,kk)=w(kk)
             op(j,2,kk)=x(kk)
122          continue
104       continue
c
          read(1,100)nop4
          nop=nop + nop4
          do 1041 j=nop1+nop2+nop3+1,nop
             read(1,102)(w(i),i=1,14),
     +                  (x(k),k=1,14)
          do 1221 kk=1,14
             op(j,1,kk)=w(kk)
             op(j,2,kk)=x(kk)
1221          continue
1041       continue
c
          close(1)
          write(5,1108)nop1,nop2,nop3,nop4,nop
1108      format(5(2x,i4))
c
c
          open(1,file='p3.dat',status='old')
          read(1,100)nop5
          nop=nop + nop5
          do 105 j=nop1+nop2+nop3+nop4+1,nop
             read(1,102)(w(i),i=1,14),
     +                  (x(k),k=1,14)
          do 123 kk=1,14
             op(j,1,kk)=w(kk)
             op(j,2,kk)=x(kk)
123          continue
105       continue
c
          read(1,100)nop6
          nop=nop + nop6
          do 1051 j=nop1+nop2+nop3+nop4+nop5+1,nop
             read(1,102)(w(i),i=1,14),
     +                  (x(k),k=1,14)
          do 1231 kk=1,14
             op(j,1,kk)=w(kk)
             op(j,2,kk)=x(kk)
1231          continue
1051       continue
c
          close(1)
          write(5,1109)nop1,nop2,nop3,nop4,nop5,nop6,nop
1109      format(7(2x,i4))
c
c
100       format(i4)
102       format(2(2x,14a1))
111       format(/)
```

```
c
c
          write(5,111)
          do 120 m=1,nop
             write(5,102)(op(m,1,i),i=1,14),
     +                   (op(m,2,k),k=1,14)
             write(*,102)(op(m,1,i),i=1,14),
     +                   (op(m,2,k),k=1,14)
120          continue

c
c
          write(5,111)
          do 1100 i=1,14
             test(i)=op(1,2,i)
                  fp(1,1,i)=op(1,1,i)
                  fp(1,2,i)=op(1,2,i)
1100         continue
c
          nxx=nop
          ns=1
c
1000      continue
          ne=0
          do 2000 ix=2,nxx
             nt=0
             do 2100 jx=1,14
             if(test(jx).ne.op(ix,1,jx)) then
                nt=nt+1
                endif
2100      continue
             if(nt.eq.0) then
                ns=ns+1
c
                ne=ne+1
                if(ne.gt.1) then
                   write(*,1003)
1003      format(1x,'ne is gt 1')
                endif
c
                do 2200 kx=1,14
                   fp(ns,1,kx)=op(ix,1,kx)
                   fp(ns,2,kx)=op(ix,2,kx)
                   test(kx)=op(ix,2,kx)
2200               continue
                       mm=0
                   do 2300 mx=1,nxx
                      if(mx.eq.ix) then
                         goto 2300
                      else
                         mm=mm+1
                         do 2400 ma=1,14
                         op(mm,1,ma)=op(mx,1,ma)
                         op(mm,2,ma)=op(mx,2,ma)
2400                     continue
                      endif
2300                  continue
                  endif
2000            continue
                nxx=nxx-1
                if(ne.ne.0) then
                    goto 1000
                    endif
c
c
          do 1220 m=1,ns
             write(5,102)(fp(m,1,i),i=1,14),
     +                    (fp(m,2,k),k=1,14)
             write(*,102)(fp(m,1,i),i=1,14),
     +                    (fp(m,2,k),k=1,14)
1220          continue
          write(*,100)ns
c
          close(5)
c
          end
```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Stephen C. Macevicz

(ii) TITLE OF INVENTION: DNA restriction site mapping

(iii) NUMBER OF SEQUENCES: 6

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Stephen C. Macevicz
(B) STREET: 21890 Rucker Drive
(C) CITY: Cupertino
(D) STATE: California
(E) COUNTRY: USA
(F) ZIP: 95014

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: 3.5 inch diskette
(B) COMPUTER: IBM compatible
(C) OPERATING SYSTEM: Windows 95
(D) SOFTWARE: Microsoft Word 5.1

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE:
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Stephen C. Macevicz
(B) REGISTRATION NUMBER: 30,285
(C) REFERENCE/DOCKET NUMBER: 1002

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: (408) 252-4140
(B) TELEFAX: (408) 252-0841

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 40 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 40

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

AATTAGCCGT ACCTGCAGCA GTGCAGAAGC TTGCGT 36

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 36 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

AAACCTCAGA ATTCCTGCAC AGCTGCGAAT CATTCG 36

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 36 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGCTCGAATG ATTCGCAGCT GTGCAGGAAT TCTGAG 36

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 36 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GTTTACGCAA GCTTCTGCAC TGCTGCAGGT ACGGCT 36

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 72 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: double
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

```
AATTAGCCGT ACCTGCAGCA GTGCAGAAGC TTGCGTAAAC CTCAGAATTC      50
CTGCACAGCT GCGAATCATT CG                                    72
```